Stochastic Models for Nonstandard, High-Dimensional Data

Chad M. Schafer, InCA Group

www.incagroup.org

Department of Statistics

Carnegie Mellon University

October 2009

The Core Collaborators

Ann B. Lee, Assistant Professor

Chad M. Schafer, Assistant Professor

Peter E. Freeman, Research Associate

Susan M. Buchman, Ph.D. Student

Joseph W. Richards, Ph.D. Student

Work Supported by NASA-AISR Grant

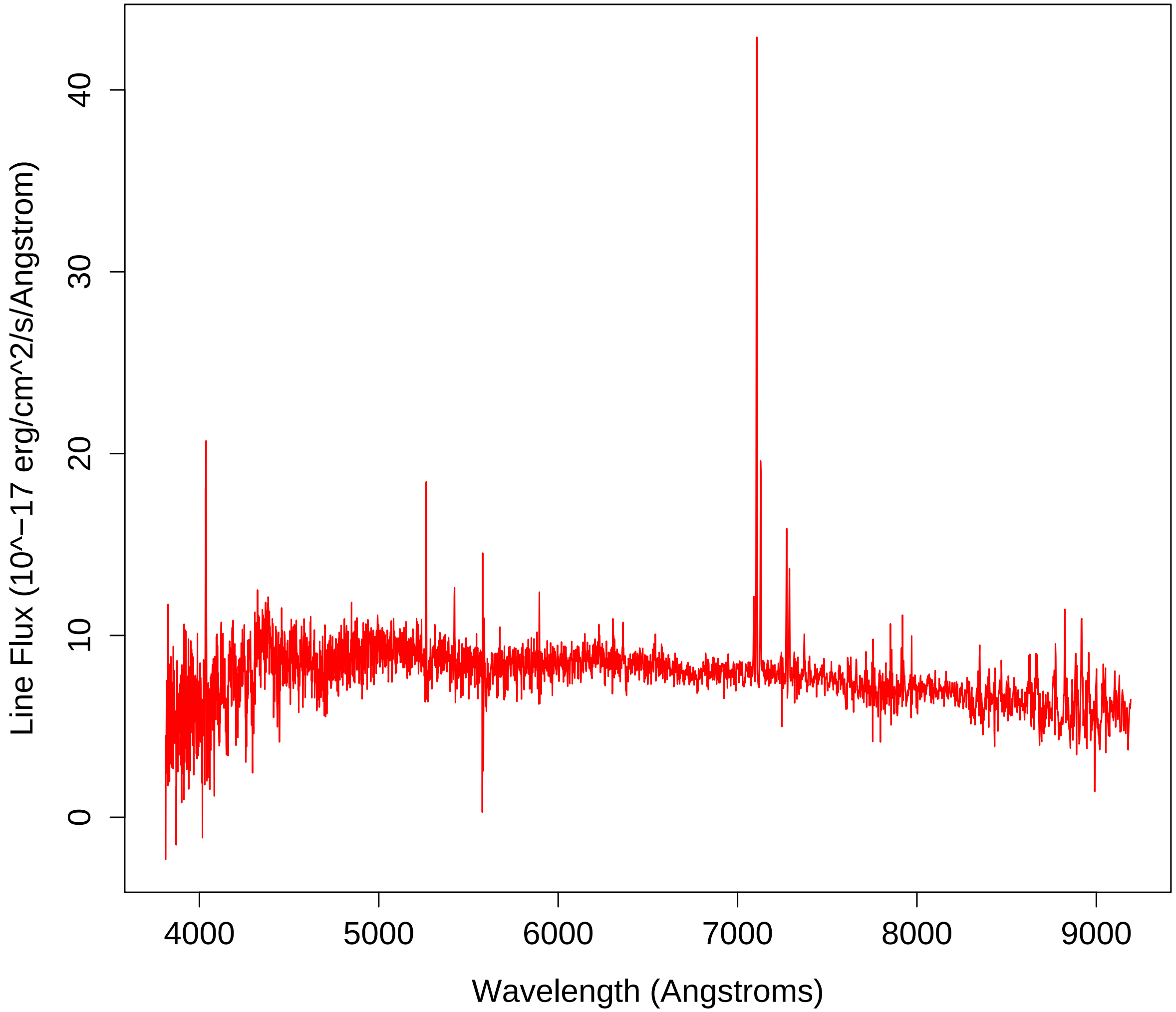
The InCA Group: `www.incagroup.org`

Motivation

Raw data are often in a form not amenable to statistical anlaysis

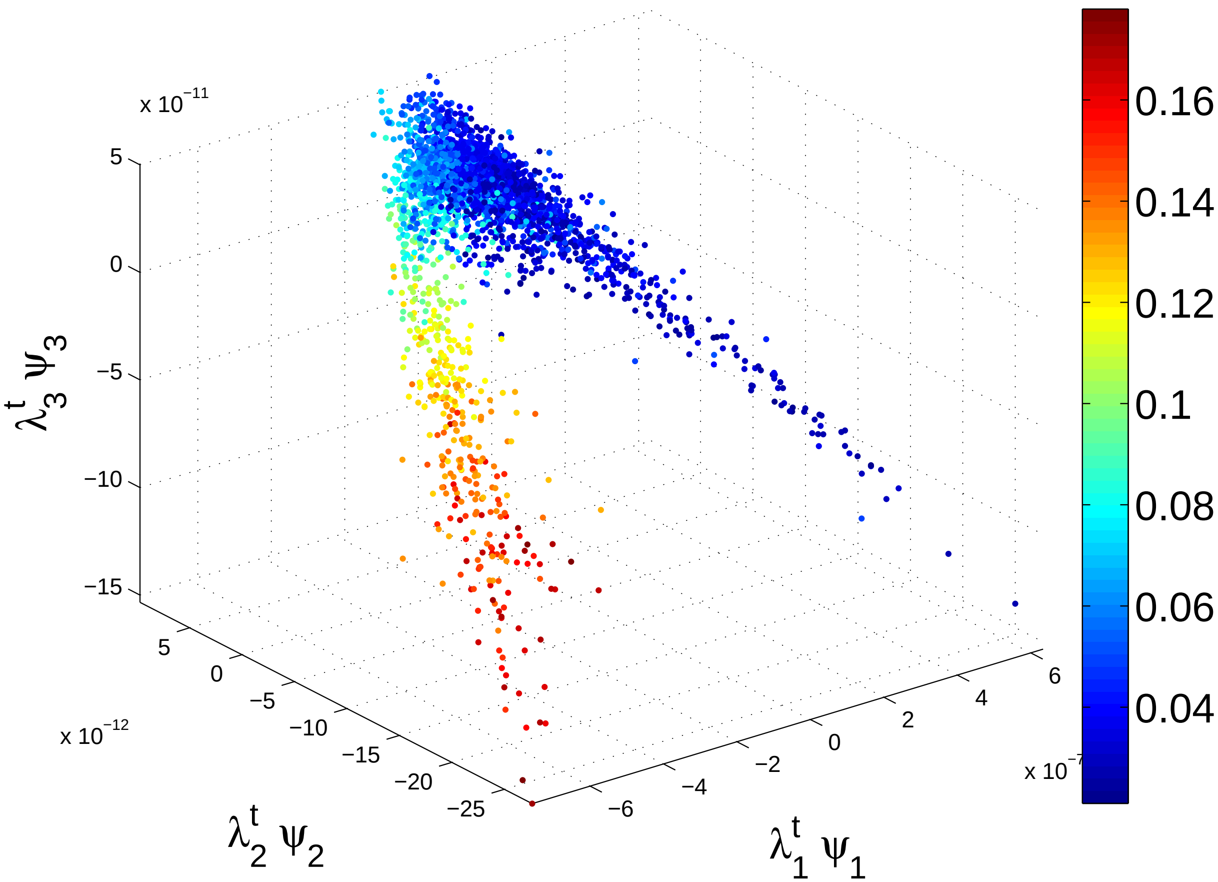
Example: Using spectra as predictors in regression

Motivation



An SDSS galaxy spectrum.

Motivation



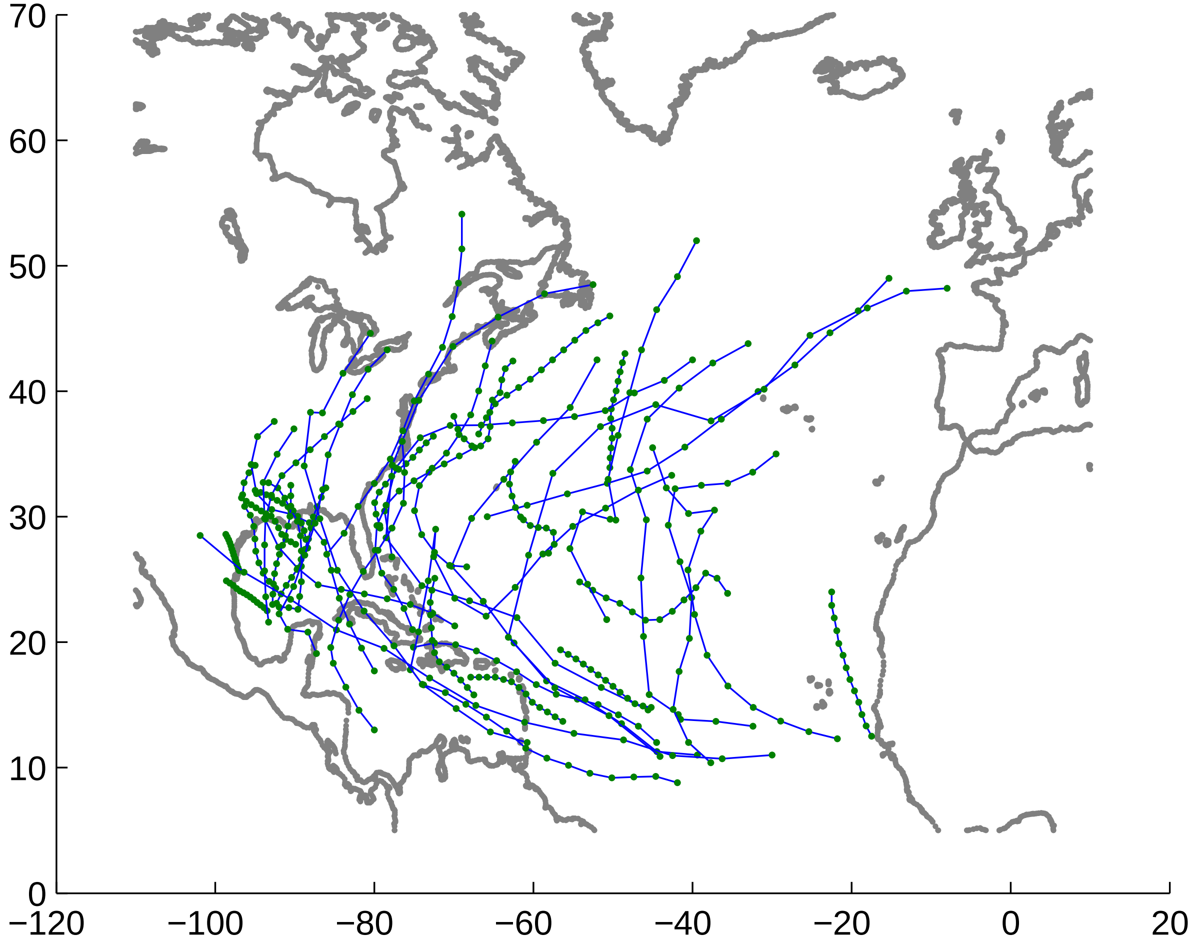
3,846 galaxy spectra, colored by redshift (Richards, Freeman, Lee, Schafer (2009a))

Motivation

Raw data are often in a form not amenable to statistical anlaysis

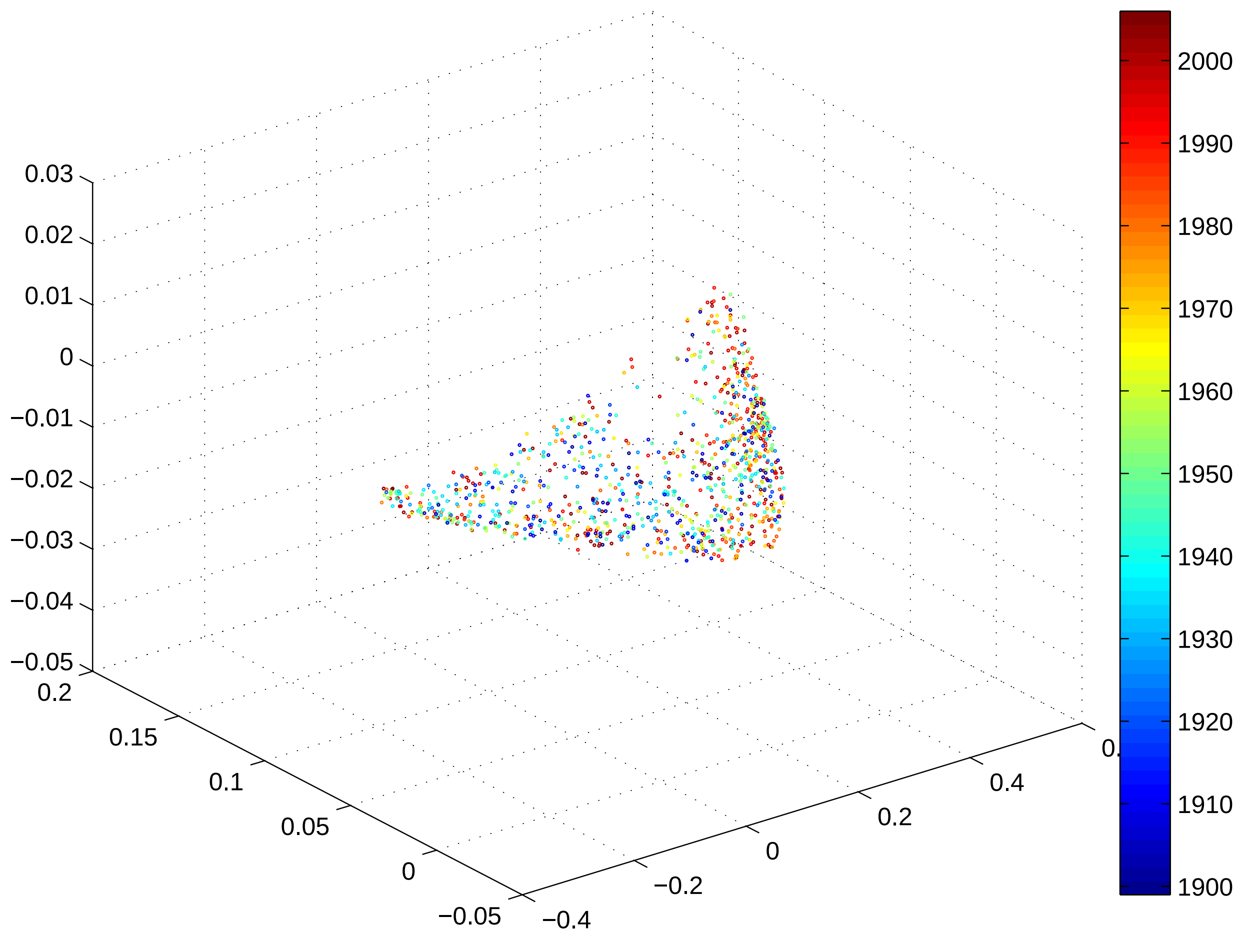
Example: Modelling the distribution of Tropical Cyclone tracks

Motivation



Tropical Cyclone (TC) Tracks (Buchman, Lee, Schafer (2009))

Motivation



1,000 TC tracks, colored by year (Buchman, Lee, and Schafer (2009))

Motivation

Reparametrize data into a new space, often of lower dimension

Data can be “nonstandard”: images, spectra, TC tracks, etc.

Location in new embedding space ideally encodes important information

Aids classification, regression, and other inference tasks

Transformations

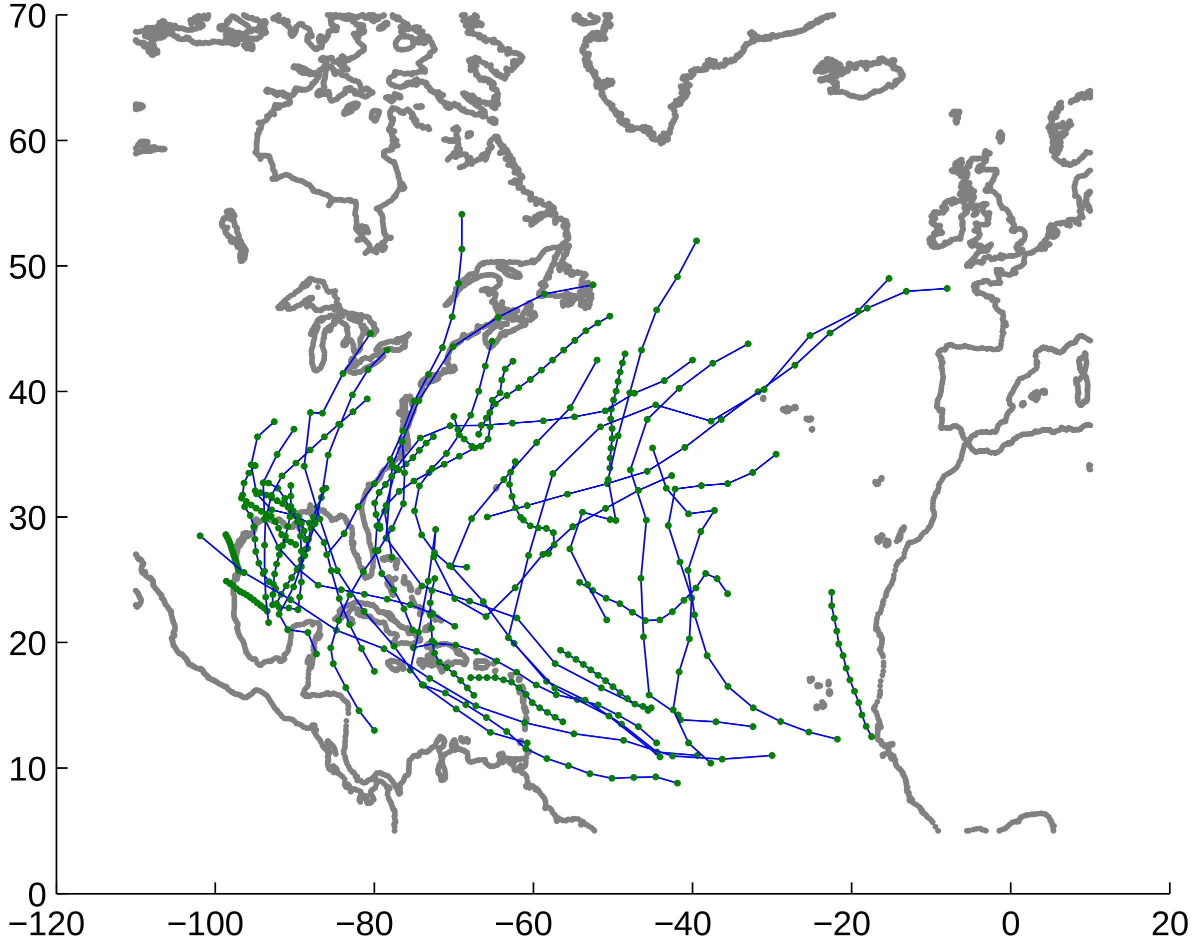
Seek embedding of data in Euclidean space that best preserves user-defined similarity/distance metric

Multidimensional Scaling

How to specify the pairwise distances?

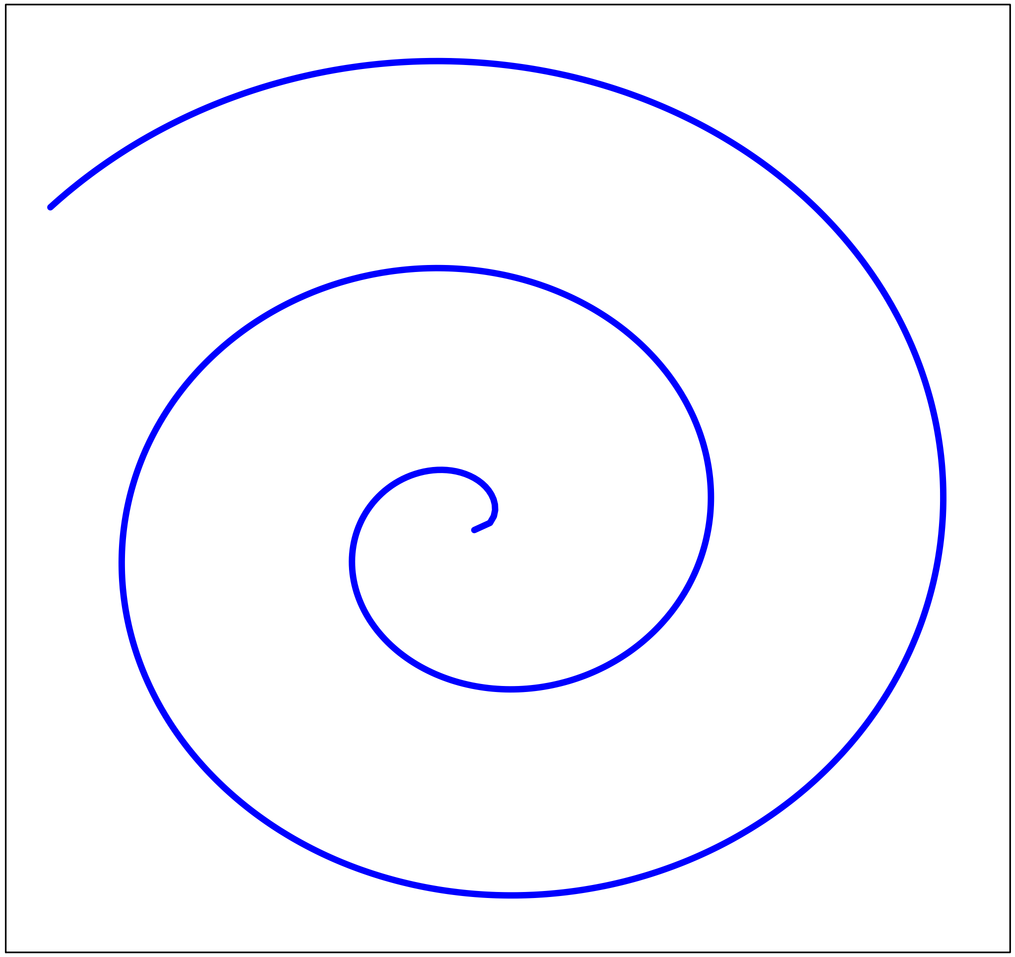
Often, we only have reliable way of judging if pairs of objects are "similar" via a local distance metric

Specifying the Distances



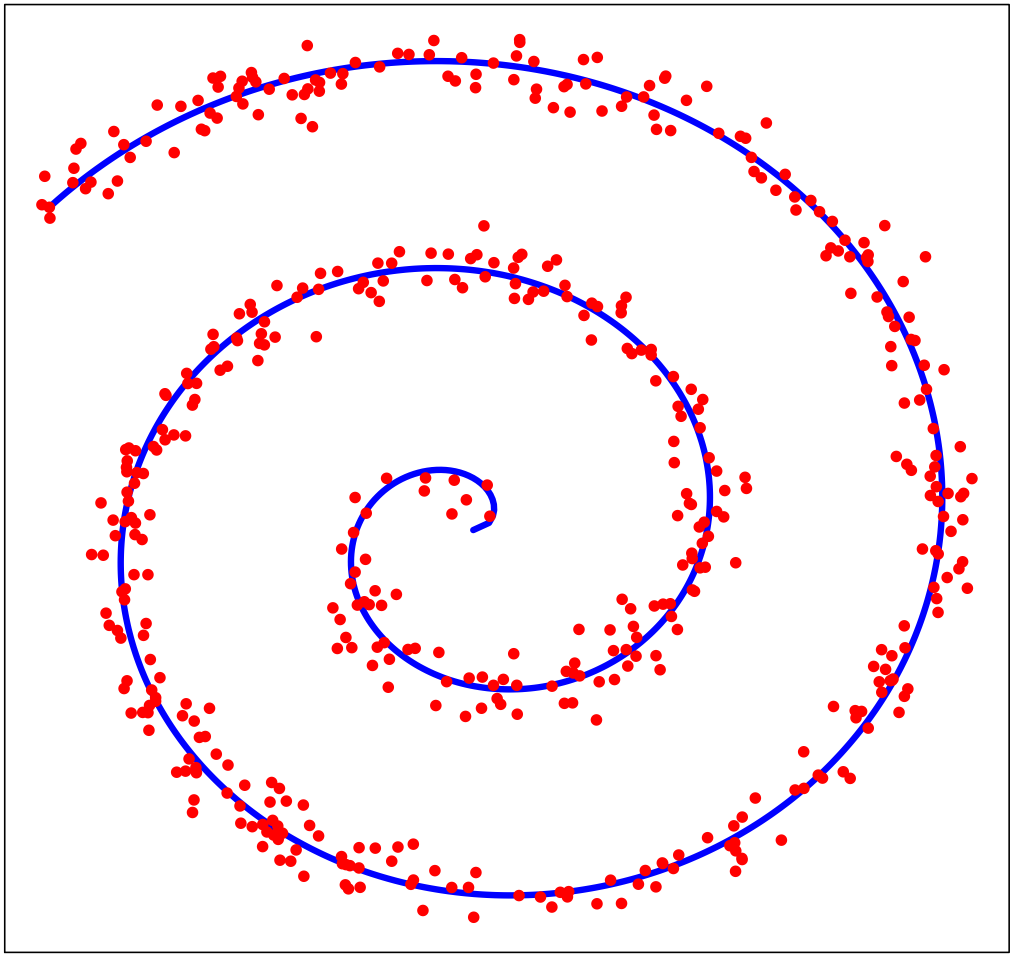
(Buchman, Lee, Schafer (2009))

Specifying the Distances



A simple, one-dimensional manifold.

Specifying the Distances



Euclidean distance good choice for local, not global, distance metric

Diffusion Distances

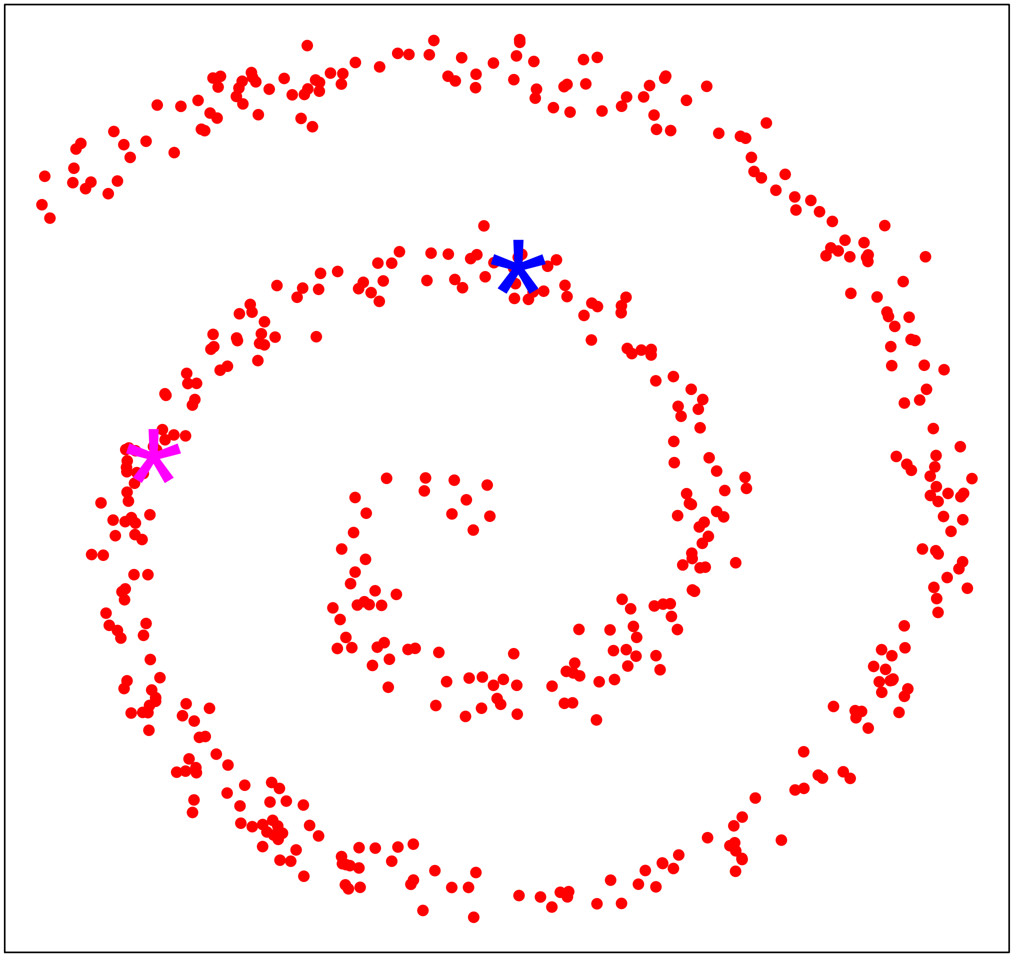
Diffusion maps are an approach to spectral connectivity analysis (Lee and Wasserman (2009))

Based on constructing fictive random walks on the data

At each “step,” can only move to “similar” data points

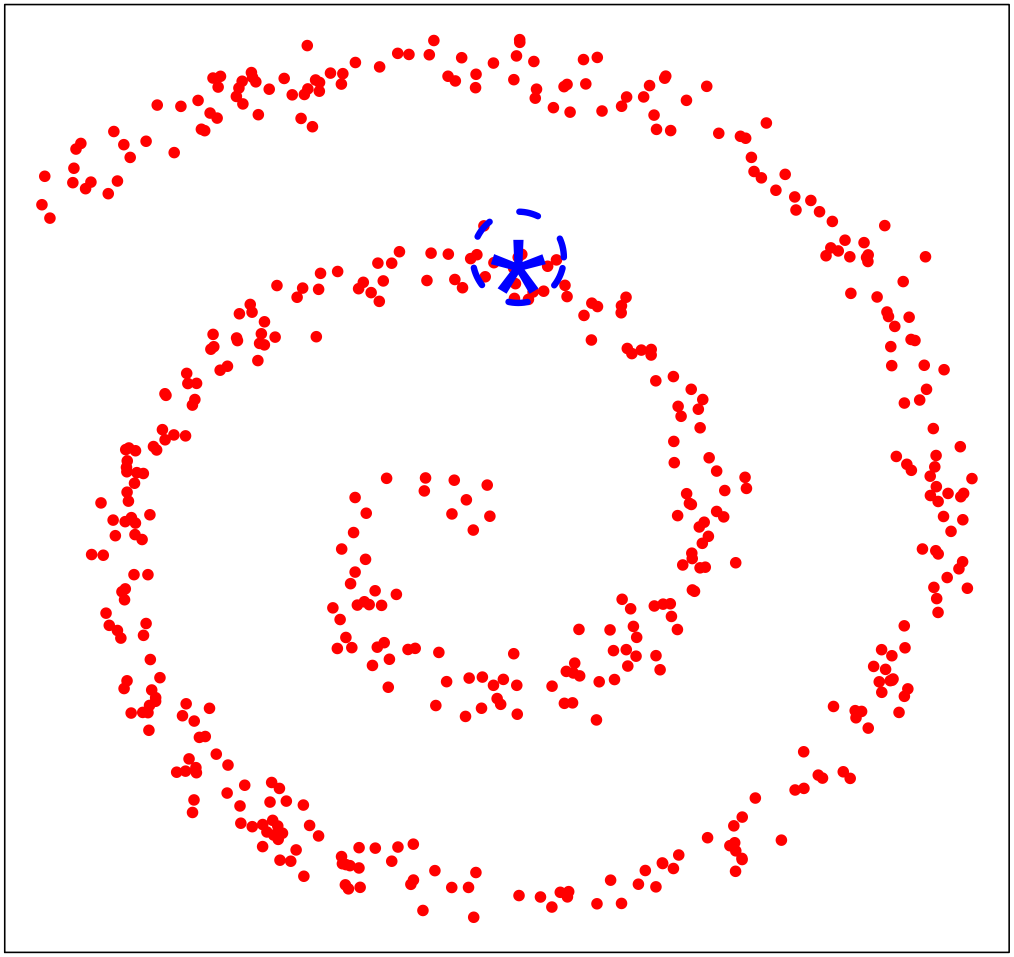
Walks starting from dissimilar data points will require many steps to “meet”

Diffusion Distances



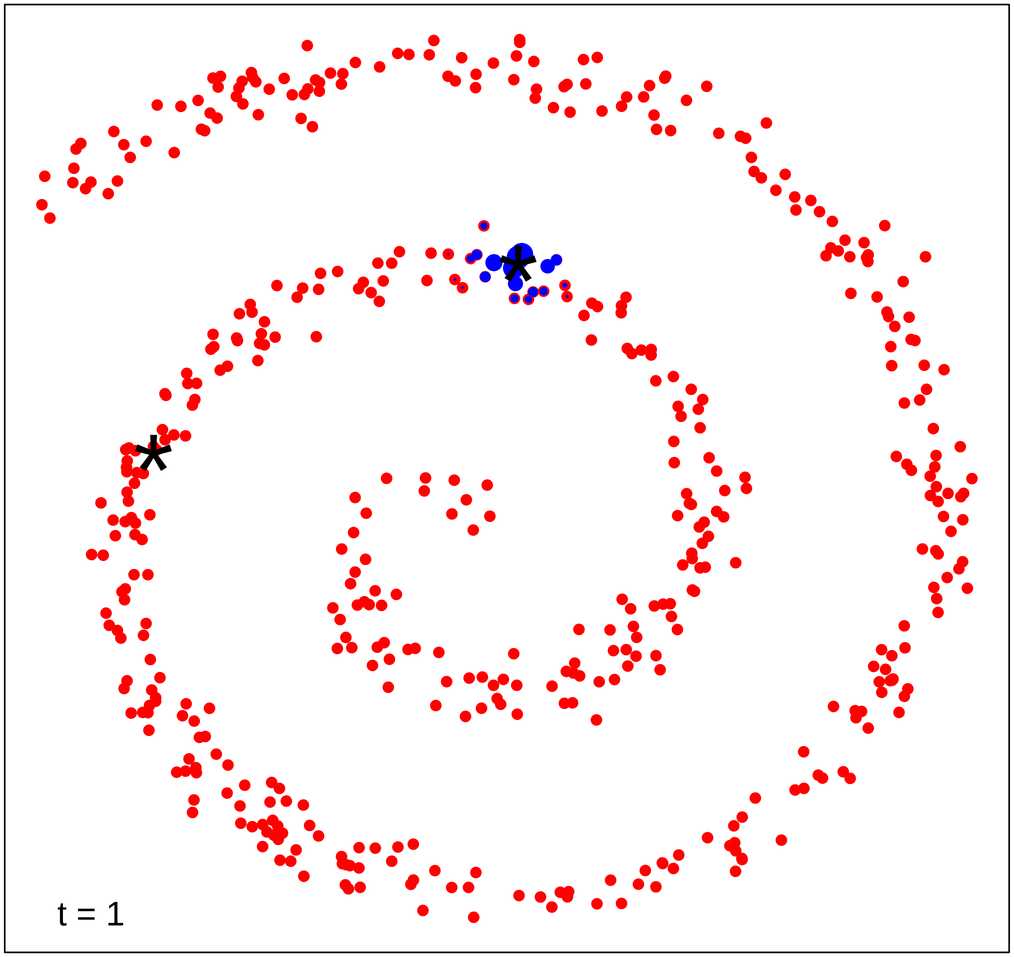
Two points on the noisy spiral

Diffusion Distances



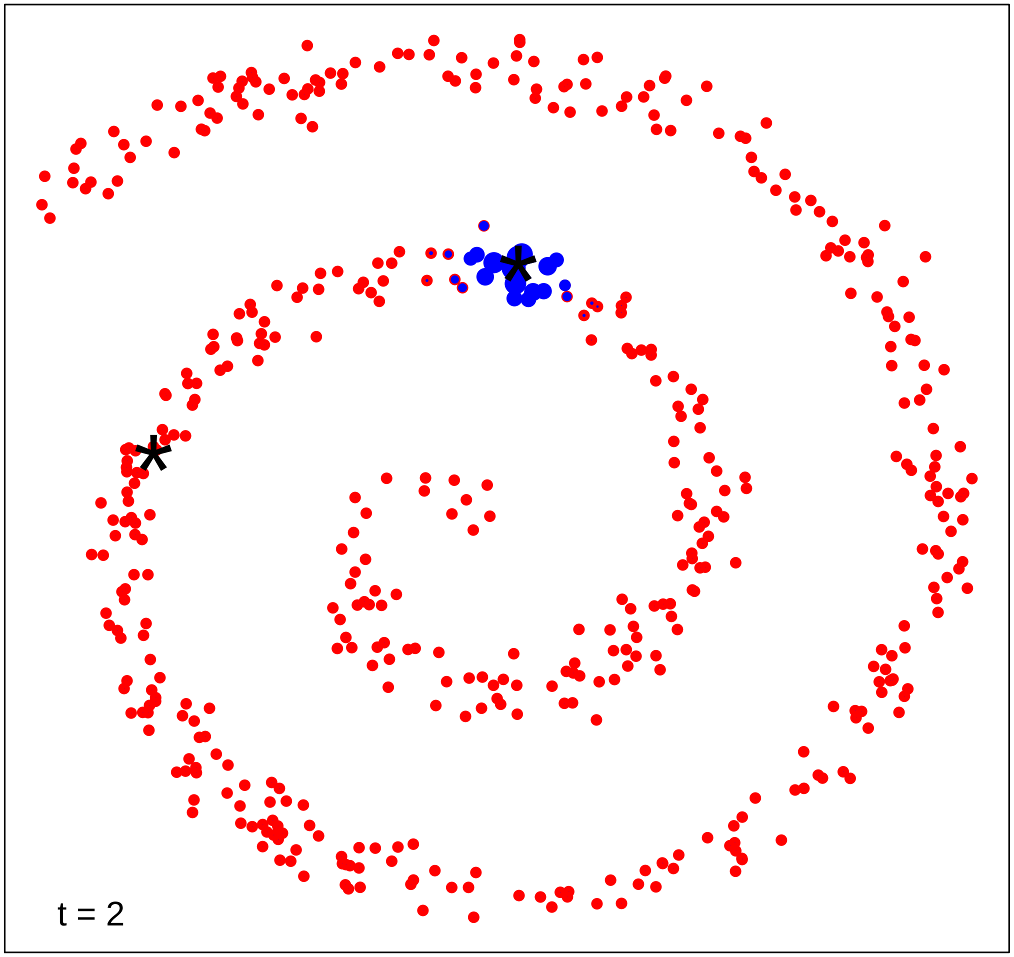
Gaussian centered on one point

Diffusion Distances



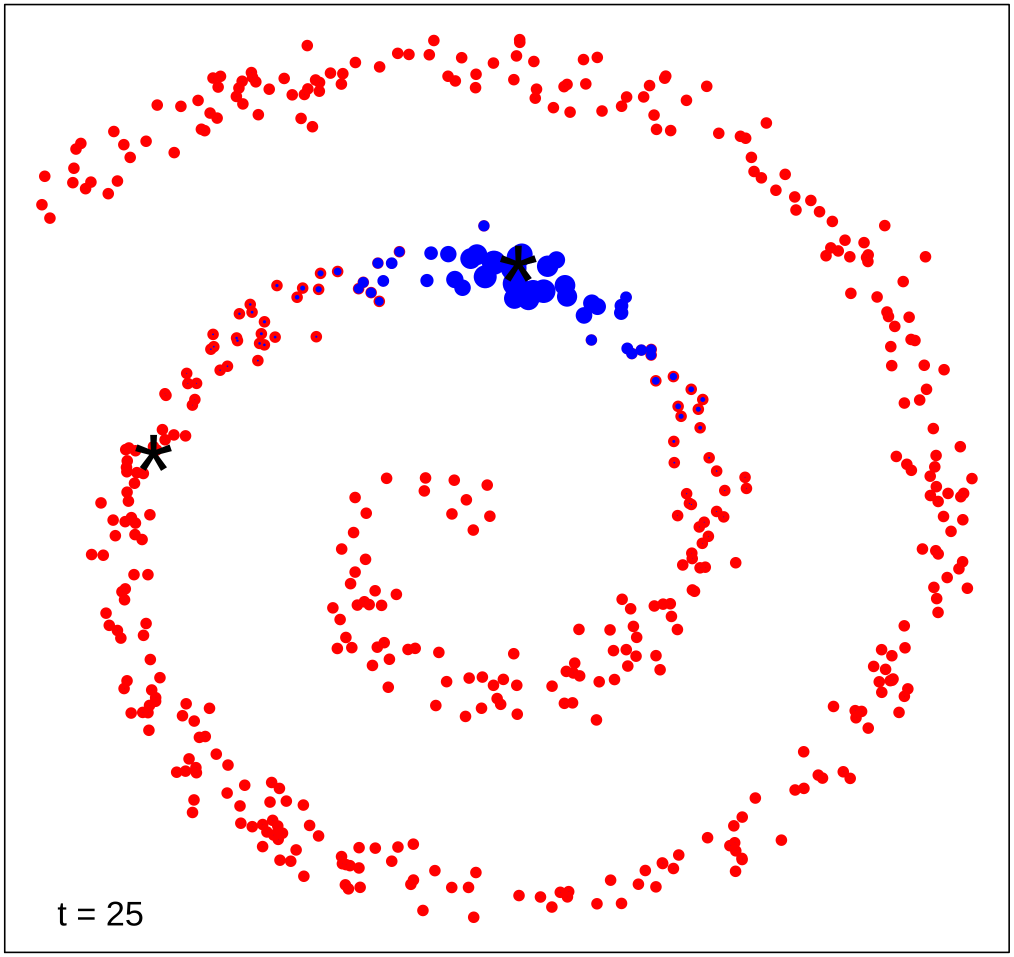
Yields distribution over points after first step

Diffusion Distances



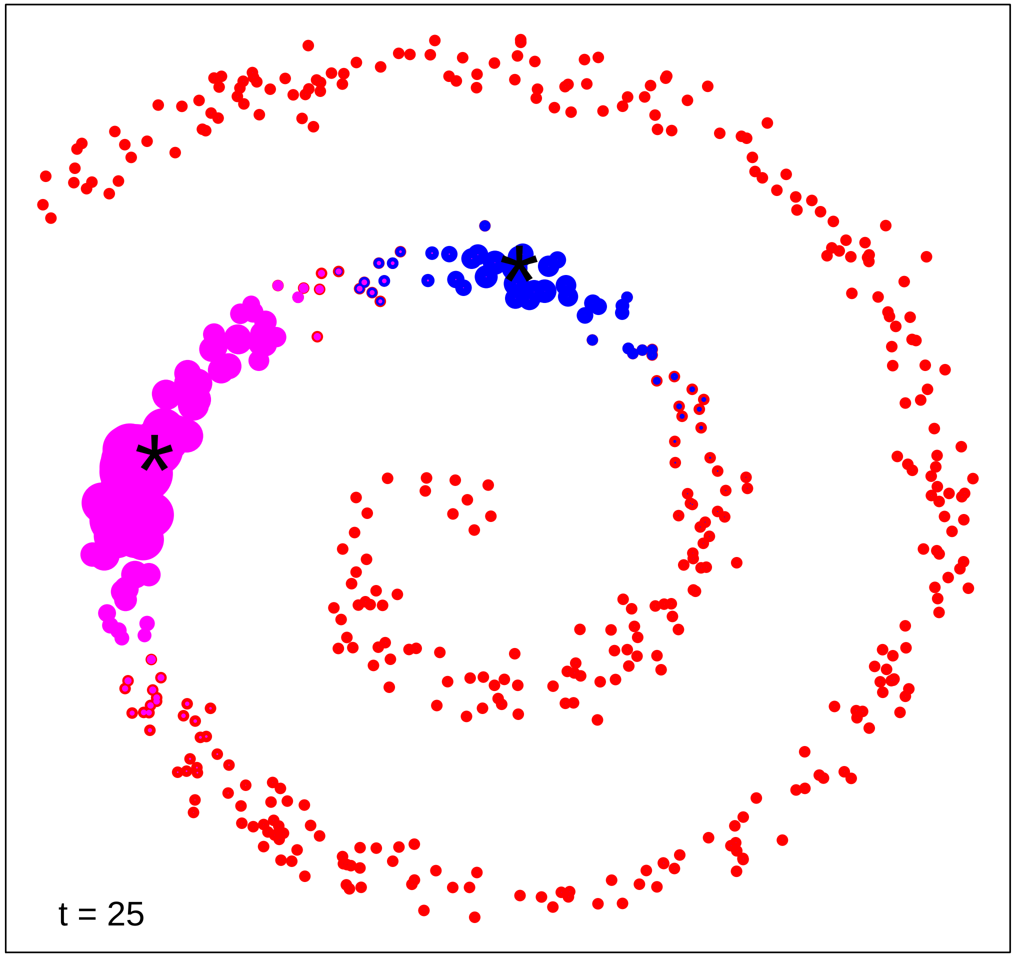
Distribution after the second step

Diffusion Distances



Distribution after the 25^{th} step

Diffusion Distances



Imagine doing for both points

Diffusion Distances

Coifman and Lafon (2006)

After t steps, a walk which begins at $\mathbf{x}$ has distribution $p_t(\mathbf{x}, \cdot)$ over $\mathcal{X}_{\text{obs}}$

As $t \to \infty$, it holds that $p_t(\mathbf{x}, \cdot) \to s(\cdot)$, where $s(\cdot)$ is the stationary distribution for the walk

Define the t-step diffusion distance between $\mathbf{x}$ and $\mathbf{y}$ as

$$D_t(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{\mathbf{z} \in \mathcal{X}_{\text{obs}}} \frac{\left(p_t(\mathbf{x}, \mathbf{z}) - p_t(\mathbf{y}, \mathbf{z})\right)^2}{s(\mathbf{z})}}$$

Diffusion Map Construction

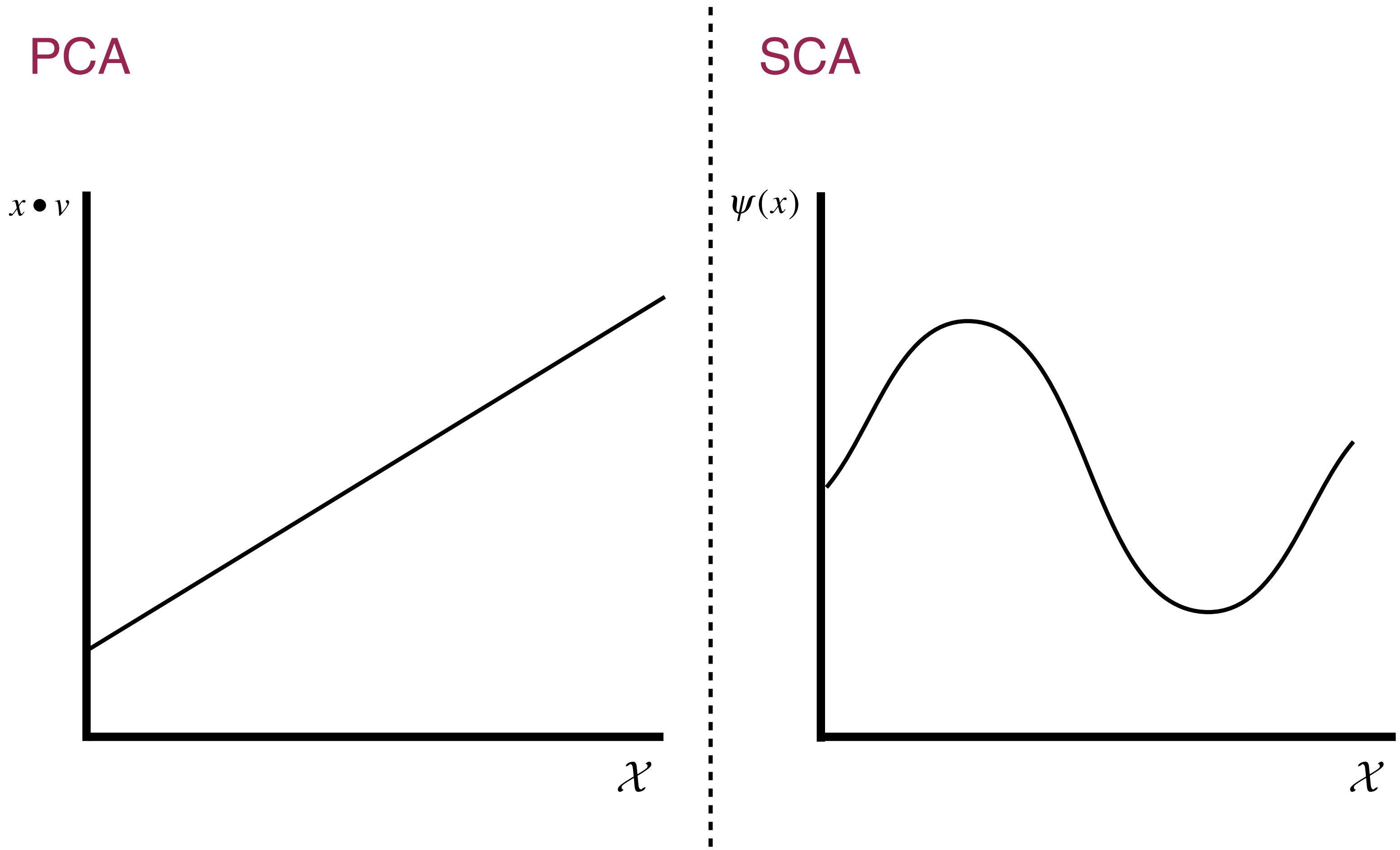
Need to specify "local" distance measure (Δ_ℓ) and neighborhood size (ϵ)

If at $\mathbf{x}$, probability next step is to $\mathbf{y}$ is proportional to

$$\exp\left(-\Delta_\ell(\mathbf{x},\mathbf{y})^2 \Big/ 4\epsilon\right),$$

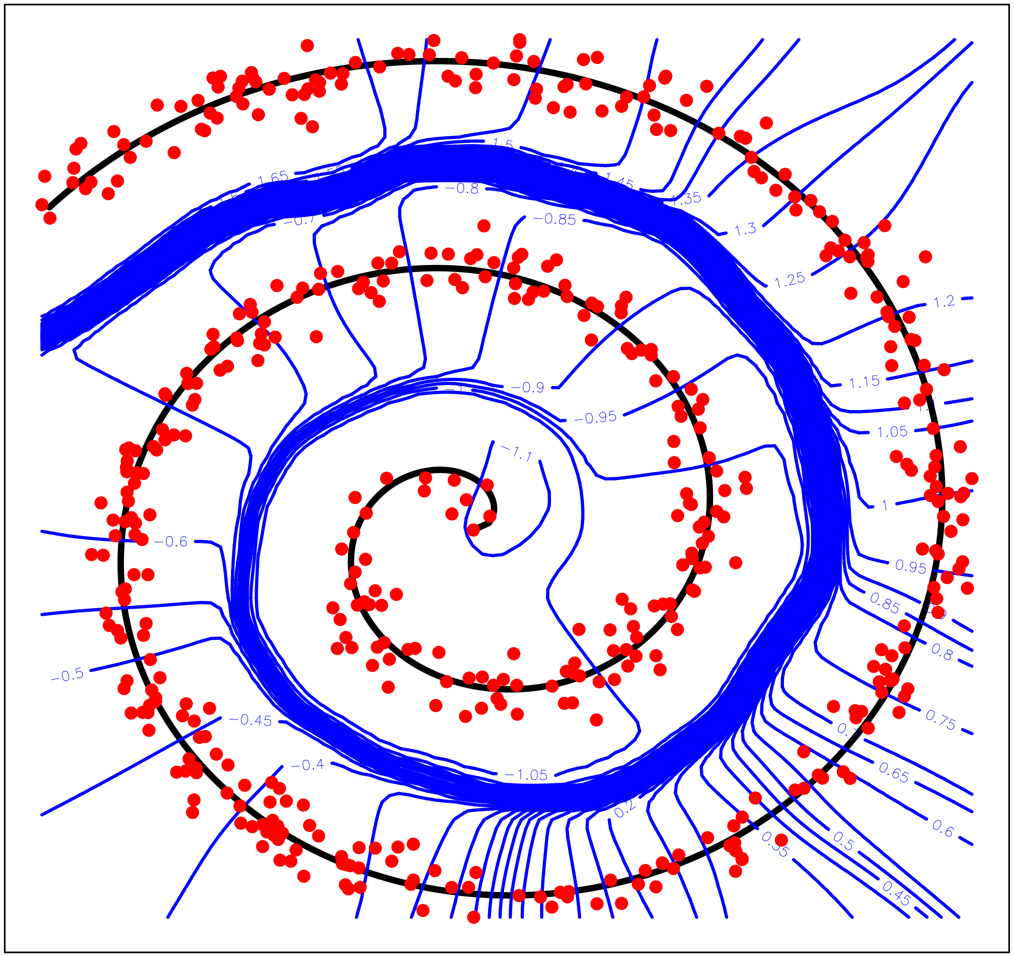
i.e., a Gaussian kernel with a "standard deviation" of $\sqrt{\epsilon}\Big/2$

Coordinate Functions



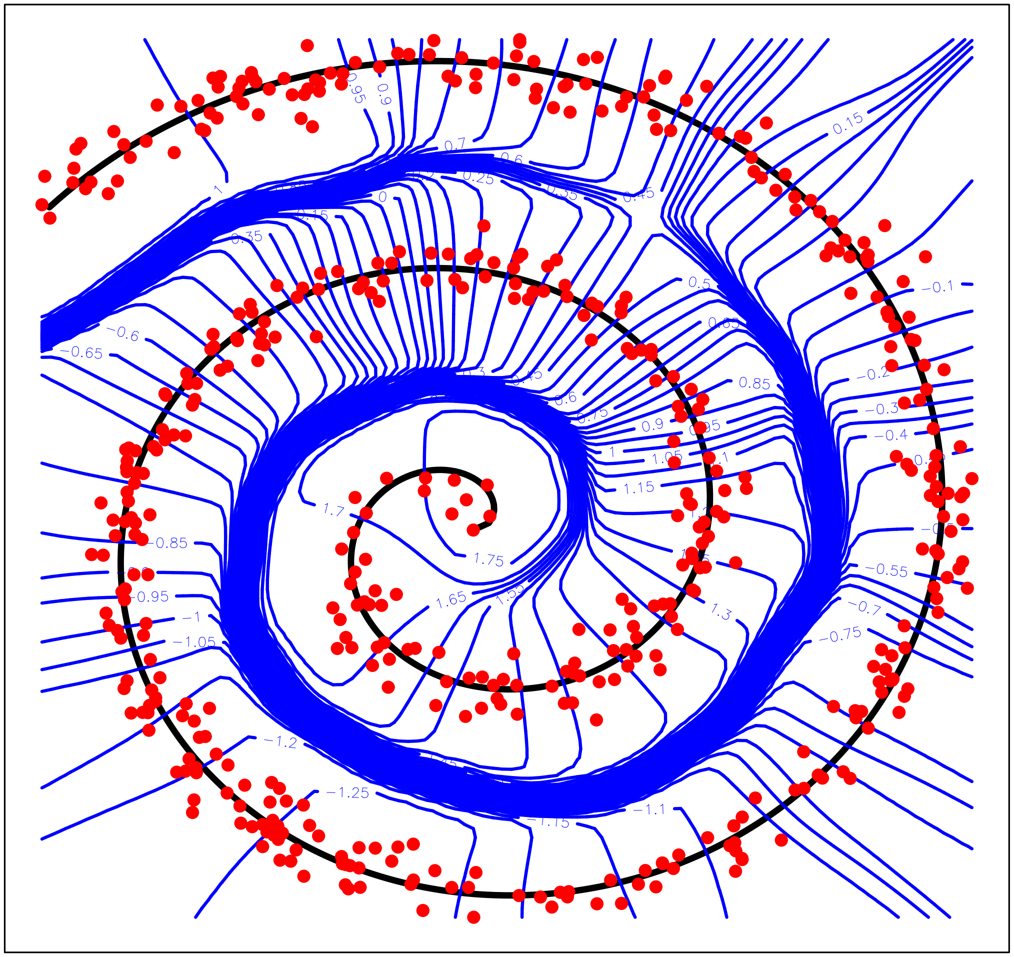
In PCA, coordinate functions are linear

Coordinate Functions



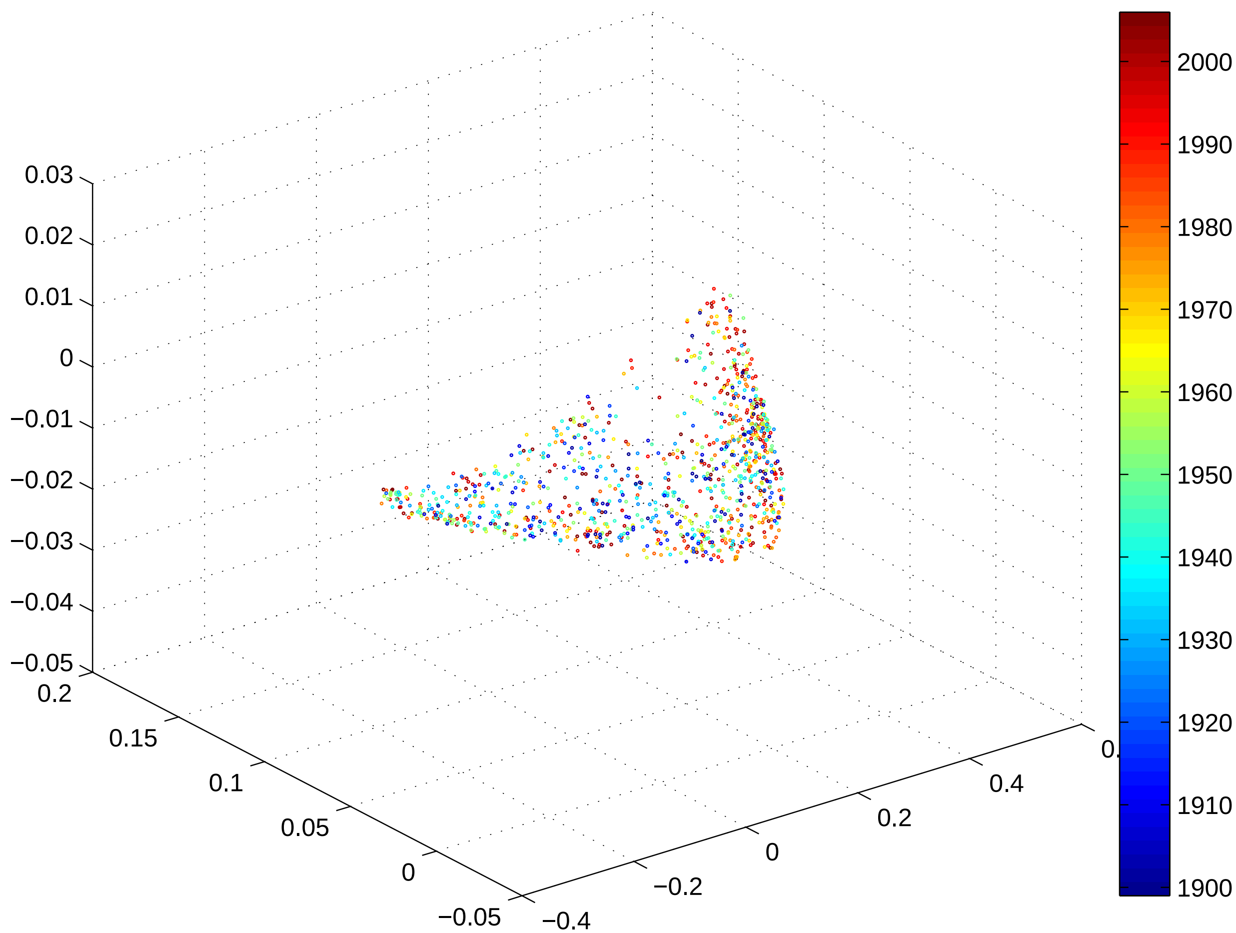
First coordinate plot for diffusion map

Coordinate Functions



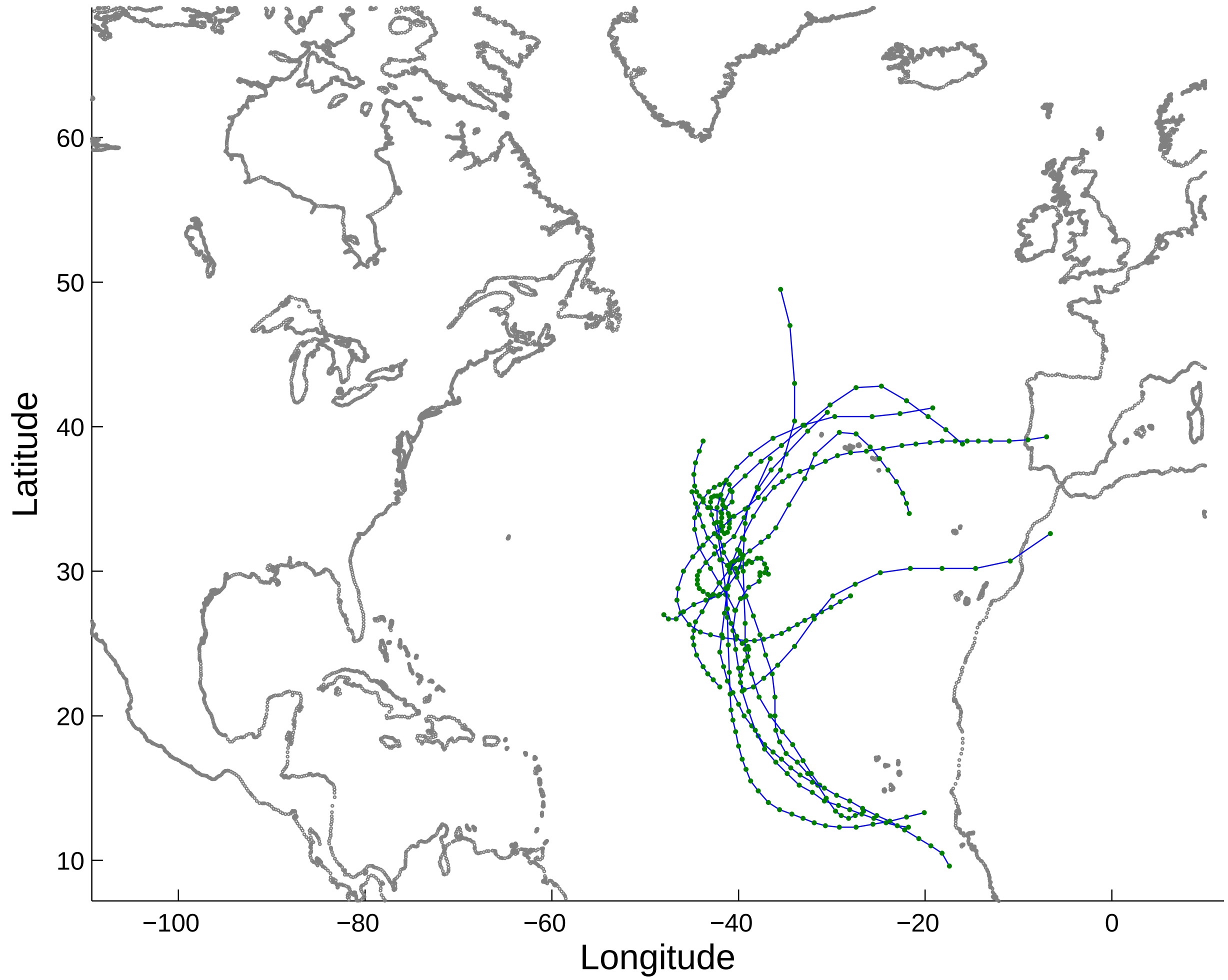
Second coordinate plot for diffusion map

Preliminary TC Results



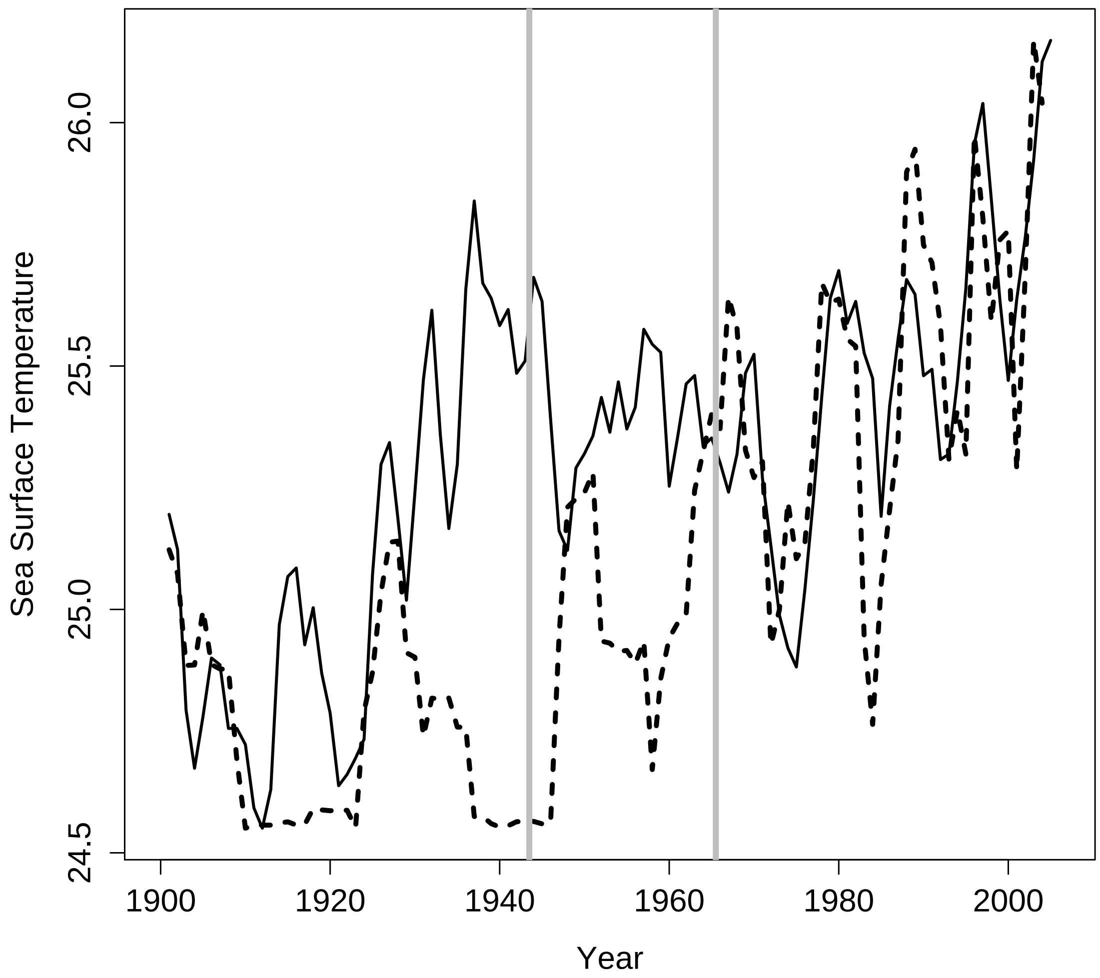
1,000 TC tracks, colored by year (Buchman, Lee, and Schafer (2009))

Preliminary TC Results



Tracks close to (0.39, 0.086, 0.0098) in diffusion space

Preliminary TC Results



Comparison of density at (0.39, 0.086, 0.0098) to SST at (30W, 15N)

Current Directions

Incorporating Covariates (climate variables)

Evolution of distribution of galaxy shapes with redshift

Comparing simulation output and real data

References

Buchman, Lee, and Schafer (2009). To appear in *Statistical Methodology.* `arXiv:0907.0199`

Coifman and Lafon (2006). *Appl. and Comput. Harmon. Anal.* **21** 5-30.

Freeman, Newman, Lee, Richards, and Schafer (2009). To appear in *MNRAS*. `arXiv:0906.0995`

Lee and Wasserman (2008). Submitted. `arXiv:0811.0121`

Richards, Freeman, Lee, and Schafer (2009a). *ApJ*. **691** 32-42.

Richards, Freeman, Lee, and Schafer (2009b). To appear in *MNRAS*. `arXiv:0905.4683`